Stoch Environ Res Risk Assess manuscript No.
(will be inserted by the editor)

Sea-surge and wind speed extremes: optimal estimation strategies for planners and engineers

Lee Fawcett · **David Walshaw**

Received: October 2014 / Accepted: date

Abstract Accurate and precise estimation of *return levels* is often a key goal of any extreme value analysis. For example, in the UK the British Standards Institution (BSI) incorporate estimates of 'once-in-50-year wind gust speeds' – or *50-year return levels* – into their design codes for new structures; similarly, the Dutch Delta Commission use estimates of the *10,000-year return level* for sea-surge to aid the construction of flood defence systems. In this paper, we briefly highlight the shortcomings of standard methods for estimating return levels, including the commonly-adopted *block maxima* and *peaks over thresholds* approach, before presenting an estimation framework which we show can substantially increase the precision of return level estimates. Our work allows explicit quantification of seasonal effects, as well as exploiting recent developments in the estimation of the *extremal index* for handling extremal clustering. From frequentist ideas, we turn to the Bayesian paradigm as a natural approach for building complex hierarchical or spatial models for extremes. Through simulations we show that the return level posterior mean does not have an exceedance probability in line with the intended encounter risk; we also argue that the Bayesian *posterior predictive value* gives the most satisfactory representation of a return level for use in practice, accounting for uncertainty in parameter estimation and future observations. Thus, where feasible, we propose a Bayesian estimation strategy for optimal return level inference.

Keywords Bayesian inference · block maxima · extremal index · extreme value theory · peaks over thresholds · return levels

L. Fawcett
School of Mathematics & Statistics, Newcastle University, Newcastle upon Tyne NE1 7RU, U.K.
Tel.: +44-191-2087228
E-mail: lee.fawcett@ncl.ac.uk

1 Background

1.1 Practical motivation

The relatively recent increase in frequency, and severity, of destructive stormy weather in the UK has stirred renewed interest in the analysis of environmental extremes, practitioners often being motivated by the estimation of the *r-year return level* for example, the sea-surge we might expect to see over-topped once, on average, every r years. Structural failure of a sea wall is possible if extreme surges are observed; estimates of the r-year return level are used to inform the design of such structures, and so the accuracy and precision of such estimates are of paramount importance. Recent work in Fawcett and Walshaw (2007; 2012) and Eastoe and Tawn (2012) revealed estimation bias for model parameters, as well as return levels, within a standard *peaks over thresholds* (POT) framework, in some cases resulting in significant under-estimation of return levels.

Estimation precision is often hampered by a lack of data on extremes; as Davison and Smith (1990) demonstrate, confidence intervals for return level estimates can be so wide that they become practically unusable. Our aim is to exploit fully any quantifiable information on temporal dependence and knowledge of seasonal variability to maximise data usage and estimation precision, whilst avoiding altogether the aforementioned problems associated with POT analyses. Working within the Bayesian framework gives the potential to facilitate these aims still further, enabling any extremal analysis to be augmented through the incorporation of prior information. Estimates of the *posterior predictive return level* can also give the practitioner a single design parameter estimate within which uncertainty in model estimation and future observations have been properly acknowledged.

1.2 Statistical modelling

Key results in extreme value theory, discussed in detail in Coles (2001, Ch. 3), point to the generalised extreme value (GEV) distribution as a model for block maxima of independent observations, with distribution function (d.f.)

$$G(y) = \begin{cases} \exp\left[-(1+\xi(y-\mu)/\varsigma)^{-1/\xi}\right], & \xi \neq 0 \\ \exp\left[-\exp\left(-(y-\mu)/\varsigma\right)\right] & \xi = 0, \end{cases} \quad (1)$$

defined on $\{y : 1+\xi(y-\mu)/\varsigma > 0\}$, where $-\infty < \mu < \infty$, $\varsigma > 0$ and $-\infty < \xi < \infty$ are parameters of location, scale and shape respectively; the case $\xi = 0$ is taken to be the limit as $\xi \to 0$. If block maxima have limiting distribution as given by (1), then an alternative characterisation of extremes, in terms of magnitudes of excess over some high threshold u, leads to the generalised Pareto distribution (GPD) with d.f.

$$H(y) = \begin{cases} 1-(1+\xi y/\sigma)^{-1/\xi}, & \xi \neq 0 \\ 1-\exp\left[-y/\sigma\right] & \xi = 0, \end{cases} \quad (2)$$

defined on $\{y : y > 0 \text{ and } (1+\xi y/\sigma) > 0\}$. The parameters of the GPD are uniquely determined by those in the GEV: specifically, the GPD scale $\sigma = \varsigma + \xi(u-\mu)$. Results in Leadbetter *et al.* (1983) show that, in the presence of short-term dependence, distributions (1) and (2) will be powered by the *extremal index* $\theta \in (0,1)$, a key parameter quantifying this dependence[1]: as $\theta \to 0$ we see increasing dependence in the extremes of the process.

The r-year return level z_r can then be obtained by inversion of $G^\theta(z_r)$ or $H^\theta(z_r)$. For example, in the case of threshold excesses, on equating to $1-r^{-1}$ this gives

$$z_r = \begin{cases} u + \sigma\xi^{-1}\left[\left(\lambda_u^{-1} w_r\right)^{-\xi} - 1\right] & \xi \neq 0 \\ u - \sigma \log\left(\lambda_u^{-1} w_r\right) & \xi = 0, \end{cases} \quad (3)$$

where $w_r = 1 - \left[1-(rn_y)^{-1}\right]^{1/\theta}$, λ_u is the rate of threshold excess and n_y is the (average) number of observations per year. An estimate of z_r, say $\hat{z}_r$, is usually obtained by replacing σ and ξ in Equation (3) with their maximum likelihood estimates $\hat{\sigma}$ and $\hat{\xi}$. A typical threshold-based analysis circumvents the estimation of θ by fitting the GPD to a set of independent *cluster peak excesses*; a filtering scheme extracts the single largest observation within a cluster of excesses of u, these clusters terminating once a run of κ consecutive sub-threshold observations is made. Thus, it is assumed that the extremes being used are independent, giving $\theta \approx 1$ in (3) and a POT analysis, as referred to Section 1.1.

To date, no general theory for non-stationary extremes has been established. As Coles (2001, Ch. 6) discusses, ignoring such non-stationarity can lead to bias in estimates of model parameters. In practice, pragmatic solutions have been proposed based on the type of non-stationarity observed. For example, trend can be incorporated through linear modelling of the GEV location parameter. More generally, the extreme value parameters can be written in the form $h(X^T\beta)$ where h is a specified function, β is a vector of parameters and X is a model vector. Smoothly varying seasonal model parameters, or a simpler seasonal piecewise approach, can also be used to account for seasonal variability (see Section 2.2, and Coles (2001, Ch. 6) for more examples; more generally, see Jonathan *et al.* (2014) for a comprehensive review). In Section 3.3 we review recent developments for modelling dependence between extremes which occurs spatially.

1.3 Illustrative applications

Figure 1 (left) represents a series of 3-hourly sea-surges collected at Newlyn, UK (1999–2001 inclusive), and (right) a section of a series of hourly gust wind speed maxima collected at Bradfield, a high altitude location in the UK (1995–2004 inclusive). These plots reveal clear seasonal variability in the wind climate at Bradfield, as well as extremal serial correlation in both datasets. Table 1 ("Block maxima") shows maximum likelihood estimates for three return levels when fitting to the set of 10 annual maximum wind speeds and the set of 36 *monthly* sea-surge maxima (we have just three years of sea-surge data so were required to use a block size smaller than the calendar year). Also shown ("Threshold excesses") are the same estimates based on a POT analysis with $\kappa = 10$ hours and $\kappa = 30$ hours for the wind speeds and sea-surges respectively ($\kappa = 30$ to allow for wave propagation time; see Coles and Tawn, 1991). *Mean excess plots* (see Coles (2001, Ch. 4)) were used to identify suitably high thresholds. To avoid issues of seasonal variability in the wind speed data, attention was restricted to extremes in the month of January wherein the largest wind speeds occur. In both analyses, the delta method (see, for example, Coles (2001, Ch. 2)) has been used to obtain standard errors for $\hat{z}_r$, although, owing to the severe asymmetry of the likelihood surface for return levels, confidence intervals have been obtained after having profiled the likelihood. The gain in precision by using a POT approach is obvious in the analysis of wind speeds. Of course, return level estimates can only be trusted if we have confidence in the fitted model from which we are extrapolating. The standard graphical diagnostics described in Coles (2001), including probability plots and quantile plots (not shown here), indicate suitable fits for both the block maxima and threshold excess analyses summarised in Table 1. In fact, further investigations revealed suitable fits of the GEV / GPD to block maxima / threshold excesses, respectively, across a range of block lengths / cluster termination intervals.

[1] Provided their "$D(u_n)$ condition" holds; informally, this condition ensures that, for large enough lags, any dependence is sufficiently negligible so as to have no effect on the limit laws for extremes.

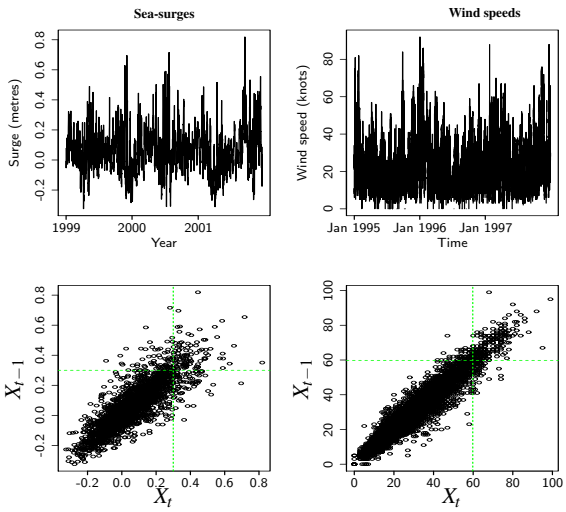

Fig. 1 Left-hand-side: Newlyn sea-surge data; right-hand-side: Bradfield wind speed data. Top: time series plots; bottom: plots of time series against series at lag 1, with thresholds. The green lines represent high thresholds used for identifying extremes.

However, Figure 2 shows the instability of return level estimates for the sea-surge data across different choices of block length / cluster termination interval. Most striking from these plots is the instability of the estimated 95% confidence upper bounds: in block maxima analyses this increases by almost 17 metres for $\hat{z}_{1000}$ when increasing the block size from one month to two months; in POT, similar changes are observed when κ increases from 10 observations to 26 observations. When a block maxima analysis and a POT analysis both indicate suitable fits, we might then appeal to estimation precision as a reason for adopting the latter. However, sensitivity of estimates to the choice of declustering interval κ (and to some degree the threshold u itself; see Scarrott and MacDonald (2012)), as illustrated here, should be noted.

1.4 Structure of this paper

The rest of this paper will be structured as follows. In Section 2 we investigate methods for increasing the precision of return level estimates by considering approaches for pressing *all* extremes into use. In particular, we consider threshold-based alternatives to POT via explicit modelling or quantification of extremal dependence, as well as using information on extremes from all seasons. Some of our recommendations here are supported with simulations. In Section 3 we then consider the Bayesian framework for return level inference. Again, the aim is to maximise data usage by properly accounting for dependence and seasonal variation. We also demonstrate the natural extension to prediction here, and present the results of a simulation study suggesting the superiority of the *posterior predictive return level* over a standard posterior summary.

		$\hat{z}_{10}$	$\hat{z}_{50}$	$\hat{z}_{1000}$
Bradfield wind speeds (knots)	Block maxima	94.9 (4.2) *(88.8, 123.8)*	102.5 (9.5) *(94.4, 223.7)*	113.2 (26.0) *(98.6, 716.5)*
	Threshold excesses	93.7 (4.3) *(87.5, 115.6)*	100.6 (6.9) *(93.7, 151.3)*	107.3 (12.2) *(98.3, 251.3)*
Newlyn sea surges (metres)	Block maxima	0.61 (0.05) *(0.53, 0.76)*	0.79 (0.11) *(0.66, 1.24)*	1.06 (0.28) *(0.80, 2.83)*
	Threshold excesses	0.87 (0.11) *(0.77, 1.57)*	0.92 (0.12) *(0.80, 2.09)*	0.97 (0.20) *(0.82, 3.38)*

Table 1 Maximum likelihood estimates of return levels. In the block maxima analyses, blocks of 1 year / 1 month were used for the wind speed / sea-surges, giving $n = 10$ / $n = 36$ extremes; in the analyses of threshold excesses, cluster peaks were identified using ($u =$ 59.8 knots, $\kappa = 10$ hours) / ($u = 0.3$ metres, $\kappa = 30$ hours) for the wind speed / sea-surges, giving $n = 33$ / $n = 39$. Standard errors are shown in parentheses, with 95% confidence intervals in italics.

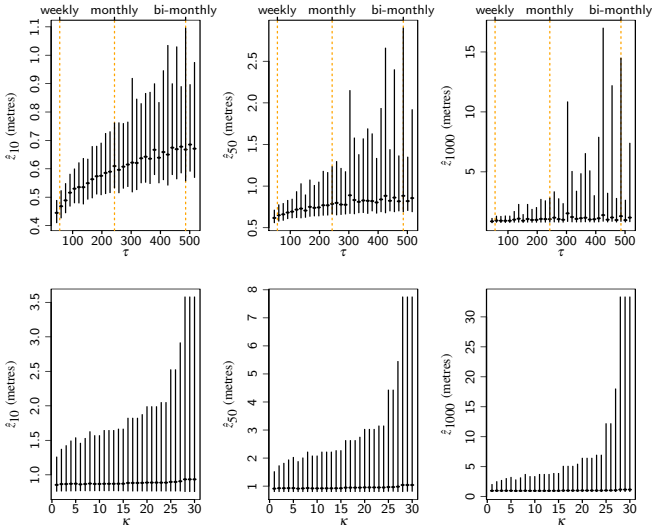

Fig. 2 Maximum likelihood estimates (points) with associated 95% profile log-likelihood confidence intervals (lines) for the 10-, 50- and 1000-year return levels for the Newlyn sea-surges. Top row: results from an analysis of block maxima with block length τ; bottom row: results from a POT analysis with declustering interval κ.

2 Increasing the precision of estimated return levels

In this Section we review some methods that have been proposed for increasing the precision of estimated return levels by *exploiting*, rather than *removing* (as in a POT analysis), any structure in the data owing to temporal dependence. In the case of the wind speed data, we also consider making use of extremes across all seasons - rather than simply the season within which the largest extremes are observed.

2.1 Serial correlation

2.1.1 Markov chain models

The POT approach for excesses over some high threshold u, as demonstrated in Section 1.3, has become standard practice in many areas of application. However, some authors

(e.g. Smith *et al.*, 1997; Fawcett and Walshaw, 2006) have explored the possibility of explicitly *modelling* within-cluster behaviour – an interesting exercise in its own right, in terms of the clustering characteristics of environmental series – but an approach which can allow the inclusion of *all* threshold excesses in the analysis. For example, based on the evidence given by plots such as those in the bottom row of Figure 1, or perhaps inspection of the partial autocorrelation function, we might assume that our series $X_1, X_2, \ldots$ forms a stationary first-order Markov chain, the stochastic properties of which being completely determined by the joint distribution of consecutive pairs. Given a model $f(x_i, x_{i+1}; \psi)$ with parameter vector ψ, it follows that the likelihood for ψ is:

$$L(\psi) = \prod_{i=1}^{n-1} f(x_i, x_{i+1}; \psi) \Bigg/ \prod_{i=2}^{n-1} f(x_i; \psi). \tag{4}$$

To model threshold excesses, the denominator in Equation (4) is replaced by the corresponding univariate densities based on a limiting model for extremes, any marginal non-stationarity being handled via modelling of the parameters within this model (as discussed in Section 1.2). Bivariate extreme value theory is invoked for contributions to the numerator, of which we give a brief summary now for threshold excesses.

Suppose $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are independent realisations of a random vector (X, Y) with joint distribution function F. For suitably high u_x and u_y, the marginals for $X - u_x$ and $Y - u_y$ both (approximately) take the form given by (2), with respective parameter sets (σ_x, ξ_x) and (σ_y, ξ_y), and with associated rates of threshold excess λ_{u_x} and λ_{u_y}, respectively. Applying

$$\tilde{X} = -\left(\log \left\{ 1 - \lambda_{u_x} \left[1 + \xi_x \left(\frac{X - u_x}{\sigma_x} \right) \right]^{-1/\xi_x} \right\} \right)^{-1}$$

to X (and similarly for Y), the variable $(\tilde{X}, \tilde{Y})$ has distribution function $\tilde{F}$ whose margins are approximately standard Fréchet for $X > u_x$ and $Y > u_y$ (see Coles, 2001, Ch. 8). It can be shown (Pickands, 1981) that the joint distribution function $G(x, y)$ for a bivariate extreme value distribution with standard Fréchet margins has the representation

$$G(x, y) = \exp\{-V(x, y)\}, \quad x, y > 0, \quad \text{where} \tag{5}$$

$$V(x, y) = 2 \int_0^1 \max(q/x, (1-q)/y)\, dW(q), \tag{6}$$

and W is a distribution function on $[0, 1]$ satisfying

$$\int_0^1 q\, dW(q) = \frac{1}{2}. \tag{7}$$

A popular choice of parametric families for G is the *logistic family*, with $V(x, y) = (x^{-1/\alpha} + y^{-1/\alpha})^{\alpha}$; here, independence and complete dependence are achieved when $\alpha = 1$ and $\alpha \to 0$ respectively. See the appendix of Fawcett and Walshaw (2012) for other choices for G. In a serial context, we would replace x / y with x_i / x_{i+1} respectively. Then contributions to the numerator in (4) can be obtained by differentiation of (5) with respect to *both* x_i and x_{i+1} if $(x_i, x_{i+1}) > u$, with appropriate censoring if one of either x_i or x_{i+1} lies sub-threshold. If $(x_i, x_{i+1}) \leq u$ then the contribution to the numerator in Equation (4) is given by the distribution function evaluated at the threshold. The marginal transformation to standard Fréchet and maximisation of the Markov chain likelihood can be performed in a single sweep, resulting in (4) being the full likelihood for both marginal and dependence parameters. Return levels can then be estimated on substitution of the estimated marginal parameters into Equation (3); an estimate of the extremal index can be obtained from the estimated dependence parameter(s) from the bivariate extreme value model used - via simulation (as in Smith (1992) or Fawcett (2005)), or via a polynomial approximation for θ (Fawcett and Walshaw, 2012).

Parametric modelling of the dependence structure requires an appropriate choice of model, as well as a suitable choice of model *order* d. Coles and Tawn (1991) demonstrate some diagnostic procedures for assessing the suitability of a first-order dependence structure ($d = 1$) relative to higher-order dependencies, but interpretation of 'simplex plots', for example, can be subjective. For $d > 1$ evaluation of the likelihood also becomes computationally expensive very quickly. Comparison of non-nested dependence models can require *ad hoc* checks of model goodness-of-fit, the interpretation of which can be subjective (e.g. Smith *et al.*, 1997). More crucially, perhaps, is the assumption of *asymptotic dependence* when using (5). Of course, standard time series models with sub-asymptotic dependence (e.g. an $AR(1)$ model) could be used instead, but graphical tools to assess the nature of the dependence (e.g. the $\bar{\chi}$ dependence measure; Coles *et al.*, 1999) can be difficult to interpret.

2.1.2 A non-parametric approach, with simulation study

Over the years there have been many publications on estimating the extremal index – for example, Leadbetter and Rootzén (1988); Smith (1992); Smith and Weissman (1994); Ancona-Navarrete and Tawn (2000); Ferro and Segers (2003); Süveges (2007); Fawcett and Walshaw (2008, 2012). Most work has focused on exploration of within cluster behaviour and the the clustering characteristics of extremes. However, our aim within the remit of this paper would be to use the extremal index to aid, and improve, return level estimation: to increase precision by using information on *all* extremes, whilst at the same time avoiding altogether the issue of cluster identification necessary in POT analyses. In fact, this is explored in Fawcett and Walshaw (2012), and simulations here reveal some promising results when specific estimators for θ are considered.

Appendix A1 summarises some common methods of extremal index estimation. Fawcett and Walshaw (2012) show that quantifying the degree of extremal dependence through the intervals estimator of Ferro and Segers (2003), and incorporating this estimate of θ in the estimation of return levels via equation (3), can more than double the estimation precision of return levels relative to estimates obtained from a standard POT analysis. However, Fawcett and Walshaw (2012) fail to assess the suitability of the other intervals estimators given in Appendix A1. We now present some results of a simulation study to assess the performance of various extremal index estimators and their ability to aid return level estimation; an extension of that in Fawcett and Walshaw (2012) but now including all of the estimators summarised in the Appendix. We also allow for processes other than those which assume asymptotic dependence. These estimators assume stationarity, and so any seasonal variability, for example, needs to be dealt with prior to estimation.

We simulate 1000 chains, each of length $10,000$, from several processes and with a range of serial correlations. Specifically, we simulate *first-order extreme value Markov chains*, as discussed in Section 2.1.1, using the (symmetric) logistic / negative logistic models, as well as the (asymmetric) bilogistic model; we simulate *max-autoregressive processes*, defined by

$$X_i = \max\left\{(1-\theta)X_{i-1}, \theta Z_i\right\}, \qquad i = 1,2,\ldots,$$

where X_0 and Z_i are standard Fréchet distributed (see Section 2.1.1); we also simulate *Gaussian AR(1) processes*, defined by

$$X_i = \psi X_{i-1} + \varepsilon_i, \qquad i = 1,2,\ldots,$$

where $\varepsilon_1, \varepsilon_2, \ldots$ are IID Normal $N(0, 1-\psi^2)$ random variables with X_0 being standard Normal. Smith (1992) discusses how the extremal index for first order extreme value Markov chains can be obtained via simulation; however, Fawcett and Walshaw (2012) exploit the deterministic relationship between θ and the parameter(s) in the bivariate extreme value model used to obtain simple polynomial forms here. The AR(1) process exhibits serial dependence but limiting extremal independence, and so here $\theta = 1$.

After marginal transformation of our chains to GPD($\sigma = 1, \xi$), maximum likelihood is used to fit the GPD to excesses above a threshold u, set at the 95% marginal quantile. Due to the threshold stability property of the GPD (see Coles, 2001, Ch. 4), these excesses will be generalised Pareto with scale $\sigma^* = \xi u + 1$ and shape ξ, and so at each replication $j = 1,\ldots,1000$ we will obtain $(\hat{\lambda}_u, \hat{\sigma}^*, \hat{\xi})^{(j)}$, $\hat{\lambda}_u$ being the observed rate of threshold excess. Using the methods in Appendix A1, at each replication j we also estimate the extremal index, giving $\hat{\theta}^{(j)}$; with the estimated marginal parameters, an estimate of the r-year return level $\hat{z}_r^{(j)}$ can then be obtained via Equation (3) (we use $n_y = 365.25 \times (24/3) = 2922$ in keeping the Newlyn sea-surge data). At each replication, the GPD is also fitted to the set of cluster peak excesses, extracted using runs declustering with various values of κ – in doing so, we can compare the standard POT approach, wherein $\theta \approx 1$, to the method which makes use of *all* threshold excesses.

Tables 2 and 3 summarise results from the simulation study for extremal index estimators and return level estimates, respectively, for $\xi = -0.4$ and certain levels of extremal dependence (other values for ξ, and other levels of extremal dependence, were also used – with similar findings obtained). Table 2 shows that for all simulated processes, there is a larger discrepancy between the sampling distribution mean and the true value for θ when using the cluster size methods than when using any of the intervals or maxima methods, and that the cluster size methods themselves are highly sensitive to the choice of cluster separation interval κ. The cluster size estimators also consistently have a higher root mean squared error (RMSE) than all the other estimators (although not shown here, similar findings were obtained for the blocks estimator of θ). Although the maxima methods require the determination of a suitable block size τ, using $\tau = \sqrt{n}$ seems to have produced reasonable estimates for the extreme value Markov chain and the max AR process. However, for these two processes the Ferro and Segers (2003) estimator and the K-gaps estimator of Süveges and Davison (2010) are superior when considering their estimated bias and RMSE; for both processes, the mean of the sampling distribution using the K-gaps estimator is closest to the true value for θ and the RMSE is smallest – although optimal values for the tuning parameter K have been used, following investigations in Süveges and Davison (2010), and this might be difficult to do in practice. There appears to be much larger bias in estimates of the extremal index for the AR(1) process than for the other two processes studied. However, as discussed in Ancona-Navarrete and Tawn (2000), the cluster size and intervals estimators are actually estimating $\theta(u)$ rather than θ, a threshold-based extremal index provided by a 'penultimate' expression for θ. In fact, Ancona-Navarrete and Tawn (2000) find that, for a marginal 95% threshold (as used here), $\theta(u) \approx 0.711$ for an AR(1) process ($\theta(u) \approx \theta$ for the other two processes used here). For a comparison of the performance of these estimators in the Bayesian framework, see Fawcett (2005).

Table 3 shows that return level estimates are less biased when using all threshold excesses, relative to a standard POT approach, regardless of the extremal index estimator used to quantify extremal dependence – and increasingly so as the return period gets larger. For all but the 10-year return level,

the RMSE is larger in the standard POT approach. For analyses using all threshold excesses, results are shown for the main contenders in terms of extremal index estimation (from Table 2), and there is little to distinguish between them – although return level estimates obtained using the K-gaps estimator have smaller bias and RMSE for all return periods considered. However, given the need to choose an appropriate block size τ for the maxima methods, and the tuning parameter K in the K-gaps method – both of which could be difficult to do in practice – we recommend using the intervals method of Ferro and Segers (2003) which provides a completely automatic solution to extremal index estimation. The results shown in Table 3 are for an extreme value Markov chain, but similar findings were also observed for the other two processes studied, and for different levels of extremal dependence.

2.2 Seasonal variability

The wind speed data observed at Bradfield exhibit clear seasonal variability, with the strongest gusts being observed in the winter months – particularly January (hence the restriction to the month of January in the analysis of Section 1.3). Experience suggests that, in the UK at least, assuming the calendar month as our seasonal unit satisfactorily reflects the changing nature of the wind climate, whilst resulting in approximate homogeneity *within* seasons. A modelling approach that identifies all gusts which are large *given the time of year* as extreme has the potential to increase estimation precision, relative to an approach using only data from a single season. Walshaw (1994) justifies using wind speed extremes from summer months in the UK: he points out that the same mechanism (an alternating sequence of anticyclones and depressions) is responsible for large wind speeds throughout the year – it is just the severity of these systems which gives rise to variations month-by-month. Such an argument supports the use of a seasonal piecewise approach for handling such variation, whereby a different model is fitted to extremes in each month. In the context of threshold models, we could follow the analysis of January wind speeds demonstrated in Section 1, but repeat the entire estimation procedure for extremes in all other months. Then, assuming independence between months, the monthly-varying GPD parameter estimates can be recombined when obtaining return level estimates by solving the following equation for $x = z_r$:

$$\prod_{m=1}^{12} H_m(x)^{n_m\theta_m} = 1 - r^{-1}, \tag{8}$$

where H_m is the GPD distribution function in month m with parameter set $(\lambda_{u_m}, \sigma_m, \xi_m)$, and θ_m / n_m are the extremal index / number of observations in month m.

Process	Estimation method		Mean	RMSE
EVMC ($\alpha = 0.5$) $\theta = 0.328$	Runs	$\kappa = 10$	0.280	0.054
		$\kappa = 30$	0.197	0.132
	Intervals	Ferro & Segers	0.340	0.012
		Süveges: MLE	0.411	0.088
		Süveges: IWLS	0.353	0.060
		K-gaps	0.324	0.004
	Maxima ($\tau = \sqrt{n}$)	Gomes	0.344	0.049
		Northrop	0.353	0.042
Max AR $\theta = 0.5$	Runs	$\kappa = 10$	0.454	0.072
		$\kappa = 30$	0.402	0.140
	Intervals	Ferro & Segers	0.501	0.056
		Süveges: MLE	0.513	0.068
		Süveges: IWLS	0.508	0.062
		K-gaps	0.501	0.009
	Maxima ($\tau = \sqrt{n}$)	Gomes	0.518	0.075
		Northrop	0.507	0.070
AR(1) ($\phi = 0.5$) $\theta = 1$	Runs	$\kappa = 10$	0.503	0.498
		$\kappa = 30$	0.234	0.766
	Intervals	Ferro & Segers	0.745	0.260
		Süveges: MLE	0.764	0.237
		Süveges: IWLS	0.782	0.241
		K-gaps	0.781	0.219
	Maxima ($\tau = \sqrt{n}$)	Gomes	0.849	0.170
		Northrop	0.814	0.194

Table 2 Sampling distribution means, and root mean squared errors (RMSE), for various estimators of the extremal index θ, and for three different types of process.

Simulated process: EVMC ($\alpha = 0.5$)			Estimated bias	RMSE
POT, using $\kappa = 10$		$r = 10$	–0.040	0.060
		$r = 50$	–0.054	0.072
		$r = 1000$	–0.078	0.079
All excesses, using various methods for estimating the extremal index	Intervals (Ferro & Segers)	$r = 10$	–0.032	0.049
		$r = 50$	–0.031	0.061
		$r = 1000$	–0.041	0.070
	Intervals (Süveges: IWLS)	$r = 10$	–0.032	0.060
		$r = 50$	–0.043	0.060
		$r = 1000$	–0.041	0.071
	Intervals (K-gaps)	$r = 10$	–0.020	0.041
		$r = 50$	–0.031	0.052
		$r = 1000$	–0.032	0.059
	Maxima (Gomes)	$r = 10$	–0.031	0.049
		$r = 50$	–0.032	0.062
		$r = 1000$	–0.043	0.073
	Maxima (Northrop)	$r = 10$	–0.032	0.054
		$r = 50$	–0.032	0.060
		$r = 1000$	–0.042	0.070

Table 3 Estimated bias and root mean squared error (RMSE) of return level estimates $\hat{z}_r$ for three return periods. Results are shown for (i) a standard POT approach to estimation, and (ii) the approach using *all* threshold excesses, accounting for extremal dependence through various methods of extremal index estimation. Here, the simulated chain was *first-order extreme value Markov*, with logistic dependence structure and $\alpha = 0.5$.

This monthly-varying GPD approach can be adapted to suit seasonal units of any size (depending on the data being analysed) although other methods for handling seasonal variability have been proposed, including the use of Fourier forms to allow the model parameters to vary continuously through

time (as demonstrated in Coles, 2001). However, most of these methods are computationally burdensome relative to the seasonal piecewise approach and, as Walshaw (1991) illustrates, can add little to return level inference in terms of accuracy and precision. Fawcett and Walshaw (2006a) also investigate the use of a conditional autoregressive structure to allow dependence between wind speed extremes in neighbouring months at Bradfield; again, they find no improvement in return level estimation by doing so. Work in Fawcett (2005) suggests significant differences in the GPD scale and shape for wind speed extremes in different months at Bradfield; often, to reduce over-fitting and where it is deemed appropriate to do so, a constant shape parameter is assumed.

2.3 Other forms of non-stationarity

As discussed so far, both our sea-surge and wind speed extremes are serially dependent, with the wind speed data also exhibiting seasonal variability. Across the time-frames studied, neither seem to display any temporal trend, although in many environmental series this departure from stationarity is an issue. A simple approach here could be to allow a linear / non-linear dependence of the extremal model parameter(s) on a time index. As discussed in Section 1.2, a dependence on other covariates can be incorporated in a similar fashion. Generally, pragmatic approaches have been developed to deal with the specific form of non-stationarity observed. For example, Chavez-Demoulin and Davison (2005) use smooth non-stationary general additive models for extremes, in which spline smoothers are incorporated into the GPD; Fawcett (2005) and Eastoe (2009) demonstrate a data pre-processing approach for dealing with seasonality and trend; Atyeo and Walshaw (2012) account for spatial dependence and temporal trend in a region-based hierarchical model for UK rainfall extremes; Jonathan and Ewans (2011) account for dependence between marginal extremes of significant wave height and wave direction / season; Coles and Walshaw (1994) propose a directional model for extreme wind speeds in the UK. For a more comprehensive review, see Jonathan *et al.* (2014).

2.4 Application to sea-surge and wind speed extremes

We now demonstrate the methods outlined in Sections 2.1 and 2.2 by application to the Newlyn sea-surges and Bradfield wind speeds. We assume stationarity in the sea-surge data, but deal with seasonal variability in the wind speed extremes observed at Bradfield by adopting the seasonal piecewise approach as discussed in Section 2.2.

Considering the Markov chain model approach outlined in Section 2.1.1, Fawcett and Walshaw (2006) provide a detailed investigation of the suitability of a first-order extreme value Markov assumption for the monthly-varying wind extremes. Plots of the χ and $\bar{\chi}$ dependence measures (see, for example, Coles, 2001, Ch. 8) suggest asymptotic dependence, providing some justification for using models from bivariate / multivariate extreme value theory for the temporal evolution of the process. Using a likelihood ratio test reveals that the bilogistic model, allowing for asymmetry in the dependence structure, gives no significant improvement over the simpler (symmetric) logistic model (see Section 2.1.1) when assuming first-order dependence only; although further graphical diagnostics suggest a second-order dependence assumption might be more suitable, Fawcett and Walshaw (2006) reveal that inferences for return levels barely change when the likelihood in (4) is extended to allow for longer-range dependencies. The estimated value of the logistic dependence parameter in each month m, α_m, can then be used to find the corresponding estimate of the extremal index θ_m via the cubic approximation derived in Fawcett and Walshaw (2012):

$$\theta \approx 0.013 - 0.092\alpha + 1.833\alpha^2 - 0.756\alpha^3. \quad (9)$$

Then, with the monthly-varying marginal GPD parameter estimates, these monthly-varying estimates of the extremal index can be used to estimate return levels on solution of Equation (8) for $x = z_r$. Estimates of the 10-, 50- and 1000-year return levels, with associated standard errors, are shown in Table 4. Standard errors for $\hat{\theta}_m$ (not shown here) have been obtained via the delta method, as have the standard errors for the estimated return levels; we have assumed that all covariances between dependence and marginal parameters are zero. Exactly the same procedure has been used to fit an appropriate Markov chain model for the Newlyn sea-surge extremes, but without the added complexity of seasonally-varying marginal and dependence parameters. Although a first-order dependence structure once again seemed adequate, the bilogistic model showed significant improvement over the logistic model for the sea-surge extremes; the polynomial approximation of the extremal index, derived in Fawcett and Walshaw (2012) as a function of the dependence parameters in the bilogistic model, was used to estimate the extremal index. Once again return level estimates, with standard errors in parentheses, are shown in Table 4.

Also shown in Table 4 are estimated return levels from analyses in which no parametric form for the dependence structure has been assumed; Ferro and Segers' intervals estimator, and the IWLS estimator of Süveges, have been used to estimate the extremal index, being our recommendations from the simulation study of Section 2.1.2 (note that both assume stationarity, which has been accounted for here in the wind speeds analysis). A block bootstrap procedure has been used to obtain the standard errors for these estimates

	Serial dependence		$\hat{z}_{10}$	$\hat{z}_{50}$	$\hat{z}_{1000}$
Bradfield wind speeds (knots)	None:	Cluster peaks	96.56 (13.53)	102.54 (22.78)	107.14 (43.05)
	Markov chain:	Logistic	88.46 (5.52)	96.07 (9.97)	107.64 (22.44)
	Non-parametric:	Ferro & Segers	88.89 (6.15)	92.88 (8.87)	105.00 (19.75)
		Süveges: IWLS	88.63 (6.50)	93.12 (8.93)	106.48 (21.18)
Newlyn sea surges (metres)	None:	Cluster peaks	0.87 (0.11)	0.92 (0.14)	0.97 (0.20)
	Markov chain:	Bilogistic	0.81 (0.07)	0.90 (0.11)	1.03 (0.18)
	Non-parametric:	Ferro & Segers	0.78 (0.06)	0.87 (0.09)	1.02 (0.16)
		Süveges: IWLS	0.78 (0.07)	0.89 (0.10)	1.03 (0.19)

Table 4 Estimates of the 10-, 50- and 1000-year return levels for the wind speeds at Bradfield and the sea-surges at Newlyn. Results from a standard POT analysis are shown, along with estimates from various approaches making use of *all* threshold excesses: accounting for dependence parametrically, using a Markov chain model, and using two non-parametric estimators for the extremal index.

(see Fawcett and Walshaw (2012) for full details). For information, and for comparison with the methods making use of all threshold excesses, we have also reported return level estimates obtained under a standard POT analysis. For the sea-surge data, these are exactly the estimates given earlier in Table 1; for the Bradfield data, the POT estimates shown in Table 4 are those obtained from a seasonal piecewise approach for dealing with monthly variations in extreme wind speeds. Here, we have made use of *reclustered excess plots* (Walshaw, 1994) to simultaneously identify monthly varying thresholds and cluster separation intervals $(u_m, \kappa_m), m = 1, \ldots, 12$.

The advantage of making use of all threshold excesses is obvious when we compare the standard errors of the estimated return levels, these being considerably smaller than those obtained from the POT analyses. In fact, we would advise the use of the non-parametric approach in practice, as this does not require the exploratory analyses of the dependence structure that the Markov chain models require. Although the standard errors shown in Table 4 are useful for highlighting the gain in precision when using all threshold excesses, as discussed throughout Section 2 we would probably rather *not* use these standard errors to construct symmetric confidence intervals. Instead, we recommend using a block bootstrap procedure, as outlined fully in Fawcett and Walshaw (2012, Section 4.3). Doing so, we construct B bootstrap replications of our process, yielding a collection of estimates $\{z_r^{(1)}, \ldots, z_r^{(B)}\}$, from which we can obtain bias-corrected, accelerated (BC_a) confidence intervals as proposed in Efron (1987). Fawcett and Walshaw (2012) show that such intervals give estimated coverage probabilities closer to the intended coverages than do the simpler percentile intervals. Implementing such a bootstrap scheme for the Bradfield wind speeds and Newlyn sea-surges gives confidence intervals for return levels that are appreciably narrower than those shown in Table 1; for example, the 95% profile-likelihood confidence interval for the 50-year sea-surge at Newlyn, obtained via the standard POT approach with $\kappa = 30$ hours, is $(0.80, 2.09)$metres (see Table 1); the corresponding 95% BC_a interval, using all threshold excesses and Ferro and Segers' intervals estimator for θ, is $(0.71, 1.02)$ metres. Similar comparisons are made when using Süveges' IWLS estimator for θ using all threshold excesses. For more details, see Fawcett and Walshaw (2012).

3 Bayesian inference for extremes

The primary aim of this paper is to find an optimal approach for estimating return levels. To this end, we have considered methods for increasing the accuracy and precision of our estimates. Working within the Bayesian framework lends further potential here. As we will demonstrate in Section 3.2, complex model structures can easily be estimated via Markov chain Monte Carlo (MCMC); specifically, we allow the sharing of information between sites and across seasons to increase the precision of our return level estimates. The natural extension to the *posterior predictive distribution* might also be useful for practitioners, the *predictive return level* giving a single point estimate incorporating uncertainty in parameter estimation *and* randomness in future observations. Although not fully realised in this paper, there is also the potential to increase estimation precision still further through the inclusion of expert-informed prior distributions.

The Bayesian paradigm was quite late to be adopted by statisticians working on extreme value theory and methods. For some general background, Coles (2001, Ch. 9) devotes a section to this topic, while Stephenson and Tawn (2004) review the literature in a paper which focuses on accounting for the three extremal types. Coles and Powell (1996) carry out a comprehensive review of the literature up to that date, and analyse wind data from a number of locations in the USA by constructing a prior for the GEV parameters based on estimates obtained at other locations. Among the other significant contributions, Coles and Tawn (1996) use expert knowledge to construct a multivariate prior for the GEV parameters, and Smith and Walshaw (2003) extend this idea to bivariate distributions for extreme rainfall at pairs of locations within a region. Smith (1999) considers predictive inference under the Bayesian and frequentist paradigms, and

Smith and Goodman (2000) and Bottolo *et al.* (2003) construct Bayesian hierarchical models for extreme values in insurance problems. Fawcett and Walshaw (2006a) model extreme wind speeds in a region of the UK using a Bayesian hierarchical model. Fawcett and Walshaw (2006) consider Bayesian inference for Markov chain models (also for extreme wind speeds) using a simulation framework similar to that used by Smith *et al.* (1997) to obtain estimates of the extremal index. More recently, Sang and Gelfand (2009, 2010) and Davison *et al.* (2012) demonstrate the use of Bayesian hierarchical models for environmental data which allow for spatial structure in the extremes.

In the absence of any prior specification for the parameters in an extremal model (e.g. the GEV or the GPD; see Section 1.2), it is possible to perform an analysis within the Bayesian framework through the use of objective priors (sometimes referred to as, quite misleadingly, 'uninformative', 'non-informative' or 'default' priors). This might also be a preferred approach if the complexity of the model makes inferences difficult or more cumbersome within a standard frequentist setting. Indeed, we discuss this in the context of the GPD (log) scale and shape parameters, and the logistic dependence parameter, in Section 3.1.1, where simple, independent, diffuse priors are suggested. However, a more thoughtful development of objective priors for extreme value models is given in Beirlant *et al.* (2004), wherein maximal data information (MDI) priors and Jeffreys' priors for the GPD are considered; similarly, Eugenia Castellanos and Cabras (2007) investigate the use of a Jeffreys' prior for the GPD. Ho (2010) and Cabras (2013) develop probability matching priors for the GPD, and Northrop and Attalides (2014) investigate posterior propriety for Jeffreys', MDI and uniform priors for the GEV and GPD.

3.1 Example: Wind speed extremes at Bradfield

3.1.1 Prior specification

In keeping with the spirit of this paper, we aim to make use of information on *all* threshold exceedances to maximise the precision of our return level estimates. Consider the likelihood in Equation (4), with parameter vector $\psi = (\eta_m, \xi_m, \alpha_m)$ for wind speed excesses over u_m in month m, $m = 1, \ldots, 12$, where

$$\eta_m = \log(\sigma_m - \xi_m u_m)$$

and ξ_m are the GPD (log) scale and shape, respectively, and α_m is the logistic dependence parameter for the first-order evolution of the process. As outlined in Section 2.2, the nature of the wind climate in the UK justifies the seasonal piecewise approach used. In the Bayesian context, the reparametrisation of the GPD scale to $(\sigma_m - \xi_m u_m)$ gives a parameter which is threshold-independent, allowing the specification of an objective prior for the scale at all threshold levels; working with the natural logarithm of this reparametrised scale retains the positivity of this parameter in the MCMC sampling scheme. In the absence of any expert prior information, then, we could specify the following independent, diffuse priors for the elements of ψ:

$$\eta_m \sim N(0, 10^4), \quad \xi_m \sim N(0, 10^2), \quad \alpha_m \sim U(0,1), \qquad (10)$$

$m = 1, 2, \ldots, 12$. We might expect such distributions to reflect our prior uncertainty about the marginal / dependence parameters and, in accord with the findings of Coles and Tawn (2005), we find that inferences barely change under order of magnitude changes to the variance specifications in (10). However, an investigation into the dependence structure of wind speed extremes at a location close to Bradfield (see Fawcett, 2005) suggests a logistic dependence parameter of around $\alpha_m \approx 1/3$ for all m. Thus, we consider independent Beta$(10, 19)$ priors for α_m, the variability of which we believe adequately reflects our knowledge about the dependence of consecutive wind speed extremes at Bradfield, including any uncertainty about differences in the dependence structure of extremes between the two locations. Similarly, from information gathered at this nearby location, we can specify the following bivariate Normal prior distributions for (η_m, ξ_m) at Bradfield:

$$(\eta_m, \xi_m) \sim N_2(\mu_m, \Sigma_m), \quad m = 1, \ldots, 12.$$

The components of μ_m are chosen to closely match our beliefs about what are the most likely values of (η_m, ξ_m) based on our study of monthly wind speeds at the nearby location. We specify values for $\mathrm{cov}(\eta_m, \xi_m)$ according to our beliefs regarding the covariances between these parameters at the nearby location, scaled (albeit rather crudely) to reflect our uncertainty about differences between monthly wind speed extremes at the two locations.

3.1.2 Bayesian sampling

After setting initial values for the elements in ψ (we use the prior means), a simple Metropolis step[2] is used to generate successive draws from the posterior distribution, giving $(\eta_m^{[j]}, \xi_m^{[j]}, \alpha_m^{[j]})$ at each iteration j, $j = 1, \ldots, 50{,}000$, in the sampler. Specifically, within each Metropolis step, a random walk procedure is used to generate candidate values for each of the parameters, the variances of the innovations being tuned to maximise the efficiency of the algorithm (achieving an overall acceptance probability of around 23%; see

[2] Details of MCMC techniques are now extensively published (Smith and Roberts (1993), for example) and so are omitted here.

Roberts *et al.*, 1997, for a discussion of desirable acceptance probabilities). Such MCMC sampling schemes can be easily implemented using the evdbayes package in R (Stephenson and Ribatet, 2014), including tuning of the acceptance probabilities and convergence diagnostics.

The bilogistic model, with dependence parameters (α_m, β_m), or indeed any of the standard models for extremal dependence, can be used in place of the logistic model. In the frequentist analysis of Section 2.3, a likelihood ratio test revealed that the bilogistic model, allowing for asymmetry in the dependence structure, gives no significant improvement over the simpler logistic model; in the Bayesian analysis, regardless of our choice of suitable (but independent) priors for α_m and β_m, the 95% credible intervals for $(\alpha_m - \beta_m)$, $m = 1, \ldots, 12$, covered zero – suggesting agreement with the frequentist analysis (the bilogistic model reduces to the symmetric logistic model when $\alpha_m = \beta_m$). Other posterior predictive checks, such as those demonstrated in Fawcett and Walshaw (2006), can be used to assess model suitability.

At each iteration j in the MCMC algorithm, the current posterior draw for the logistic dependence parameter $\alpha_m^{[j]}$ is used to obtain a posterior draw for the extremal index via the cubic approximation in Equation (9), giving $\theta_m^{[j]}$. Then, a corresponding draw from the posterior for various return levels $z_r^{[j]}$ can be obtained on solution of Equation (8) for $x = z_r^{[j]}$, after substitution of σ_m, ξ_m and θ_m with $(e^{\eta_m^{[j]}} + \xi_m^{[j]} u_m)$, $\xi_m^{[j]}$ and $\theta_m^{[j]}$, respectively; λ_{u_m} is fixed at the observed proportion of exceedances of u_m in each month m. The MCMC sample paths (not shown here) showed rapid convergence to their apparent stationary distributions, with good mixing properties (more formal convergence monitoring diagnostics are available – see, for example, Brooks and Gelman, 1998). After removal of the burn-in period (the first 2000 MCMC draws), we are left with $S = 48{,}000$ posterior draws on which to make inferences. Table 5 ("Standard analysis") shows posterior summaries for the 10-, 50- and 1000-year return levels for wind speeds at Bradfield after the removal of the burn-in period. Relative to using the uninformative priors in (10) (results not shown), we observe smaller posterior standard deviations; notice also that these posterior standard deviations are smaller than the estimated standard errors obtained in the frequentist analysis of Section 2.3 (see Table 4). Credible intervals in the Bayesian context (see Table 5) are also more readily available, obtained by direct reference to the posterior draws for z_r.

3.1.3 Predictive inference

Suppose we assume the same marginal and dependence structure for future extremes Y of our monthly wind speed processes at Bradfield. Allowing for uncertainty in parameter estimation and future observations, we can write

$$\Pr\{Y \leq y|x\} = \int_{\Psi} \Pr\{Y \leq y|\psi\}\, \pi(\psi|x) d\psi \tag{11}$$

for the *predictive distribution* of our wind speed extremes, where x represents past observations. Solving

$$\Pr\left\{Y \leq z_{r,\mathsf{pred}}|x\right\} = 1 - r^{-1} \tag{12}$$

for $z_{r,\mathsf{pred}}$ therefore gives an estimate of the r-year return level that incorporates uncertainty due to model estimation. Although (11) is analytically intractable, it can be approximated since we have estimated the posterior distribution using MCMC. Regarding our sample $\psi^{(1)}, \ldots, \psi^{(S)}$ as realisations from the stationary distribution $\pi(\psi|x)$, we have

$$\Pr\left\{Y \leq z_{r,\mathsf{pred}}|x\right\} \approx \frac{1}{S}\sum_{j=1}^{S} \Pr\left\{Y \leq z_{r,\mathsf{pred}}|\psi^{[j]}\right\}, \tag{13}$$

which we can set equal to $1 - r^{-1}$ and solve for $z_{r,\mathsf{pred}}$ using a numerical solver. These values are shown in Table 5 for $r = 10$, 50 and 1000. Figure 3 compares predictive and

	z_{10}	z_{50}	z_{1000}
Standard analysis	96.21 (2.38)	103.14 (4.51)	113.55 (12.16)
	(91.69, 102.80)	(96.61, 115.92)	(104.57, 149.31)
	100.71	*111.96*	*144.94*
Hierarchical model	96.89 (0.98)	103.46 (1.33)	128.13 (2.69)
	(94.96, 98.85)	(94.11, 116.06)	(117.62, 140.31)
	104.39	*113.09*	*147.34*

Table 5 Posterior means (standard deviations) and 95% credible intervals in parentheses, for the 10-, 50- and 1000-year return levels from Bayesian analyses of the Bradfield wind speed extremes. Shown in italics are estimates of the corresponding predictive return levels. Units in knots.

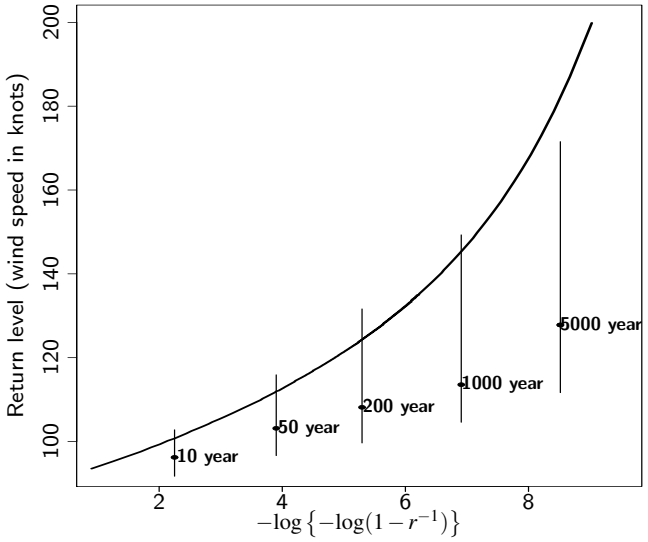

Fig. 3 Predictive return level curve (bold line) for Bradfield. Also shown, for comparison, are posterior means for some standard return levels with their 95% credibility bands.

estimative return levels across a range of values of r, showing that, for very long-range return periods, even designing a structure to the upper end-point of the Bayesian 95% credible interval might result in under-protection, relative to estimates obtained in the predictive analysis.

3.1.4 Non-parametric approaches for serial dependence

In the earlier frequentist analyses, we advocated the use of non-parametric estimators (e.g. Ferro & Segers' intervals estimator) for the extremal index rather than a Markov chain model as used in this Section. In the absence of a likelihood for the extremal index, such non-parametric methods are difficult to implement within a Bayesian sampling scheme. Ferro and Segers (2003) do propose a maximum likelihood estimator for the extremal index based on their inter-arrival times methodology. However, the model, based on a mixture distribution, one component of which is an exponential distribution with rate θ, assigns all of the inter-exceedance times to the exponential component as $n \to \infty$ (where n is the length of the process), a feature illustrated when using the associated likelihood as an ingredient in Bayesian inference for θ in Fawcett (2005): the effect of using this likelihood is a posterior distribution for θ that converges to a point mass at 1, regardless of the strength of serial correlation present.

Süveges (2007) also suggests a likelihood for θ (the corresponding maximum likelihood estimator is demonstrated in the simulation study of Section 2.1.2 of this paper); however, Table 2 reveals substantial bias when the underlying process is an extreme value Markov chain. The K-gaps estimator of Süveges and Davison (2010) is likelihood-based, and as we show in the simulation study of Section 2.1.2 it performs well when K is chosen optimally. Indeed, since the first-order extreme value Markov chain assumption, with logistic dependence, seems reasonable for our monthly wind speeds data, this could have been tried here; however, more generally it might be difficult to choose a value for K which lends optimal performance to this estimator of the extremal index. Fawcett and Walshaw (2008) demonstrate the use of a GEV likelihood which incorporates θ, proposed by Ancona-Navarrete and Tawn (2000), as an ingredient for Bayesian inference for θ, although this approach is sensitive to the block size τ that must be chosen. The semi-parametric estimator of Northrop (2012) is also based on a likelihood and so is an additional possibility in this context, although once again a tuning parameter (again the block size τ) must be chosen carefully. Thus, for Bayesian inference, we recommend using a suitable parametric form for the dependence structure in the extremes, as demonstrate in Sections 3.1.1–3.1.3.

3.2 Spatial considerations

In Section 3.1 we demonstrated the advantages of a Bayesian approach to return level inference through a rather basic application to the wind speed data at Bradfield. Even a rather crude attempt to incorporate prior knowledge into the analysis resulted in estimates of posterior variability that were substantially smaller than the asymptotic standard errors in the corresponding frequentist analysis. Prediction is also handled neatly within the Bayesian framework, as illustrated in Section 3.1.3 – estimates of predictive return levels are potentially appealing to practitioners, as they account for uncertainty due to model estimation *and* uncertainty in future observations. Another advantage of working within the Bayesian framework is the relative ease with which we can build more complex, and potentially realistic, model structures, as we now demonstrate. In the following application, return level estimation precision is increased still further.

Fawcett and Walshaw (2006a) develop a hierarchical model for extreme wind speeds observed at 12 locations in central / eastern England (Bradfield, as used throughout this paper, being one of these sites). In an attempt to share information across sites and seasons, they specify the following model structure for GPD scale and shape parameters, and the logistic dependence parameter, as used throughout Section 3.1:

$$\begin{aligned} \eta_{m,s} &= \gamma_\eta^{(m)} + \varepsilon_\eta^{(s)}, \\ \xi_{m,s} &= \gamma_\xi^{(m)} + \varepsilon_\xi^{(s)} \quad \text{and} \\ \alpha_s &= \varepsilon_\alpha^{(s)}, \end{aligned}$$

where, generically, γ and ε represent seasonal and site effects respectively, $m = 1, \ldots, 12$ being an indicator of season (month), and $s = 1, \ldots, 12$ being an indicator of site. All random effects for $\eta_{m,s}$ and $\xi_{m,s}$ were taken to be normally and independently distributed; the means and variances of the random effects distributions were given distributions that were thought to reasonably reflect prior ignorance about the seasonal and site effects, whilst retaining conjugacy wherever possible to simplify the MCMC sampling scheme. The logistic dependence parameter α was allowed to vary by site only (an *a priori* assumption justified by the nature of the wind climate across seasons within the UK; the analysis in Section 3.1 also revealed similarity in α_m across all months m), but a $U(0,1)$ prior was used for $\varepsilon_\alpha^{(s)}$ to reflect prior ignorance about the dependence structure for wind speed extremes for each site as a whole (of course, more informative priors, as specified in Section 3.1.1 for the Bradfield wind speeds, could have been used). Where conjugacy facilitated specification of full conditional distributions, Gibbs sampling was used (i.e. to obtain draws from the posterior distributions of the parameters in the random effects distributions); a Metropolis step, as discussed in Section 3.1.2,

was used elsewhere. See Appendix A2 for more details, including the full conditional distributions used in the Gibbs sampler.

Posterior summaries of return levels, at Bradfield, are shown in Table 5 ("Hierarchical model"). The effect of sharing information on extremes at other sites can be seen in the reduction of posterior variability relative to the standard Bayesian analysis (which uses information at Bradfield only, although information from a neighbouring site *is* used to aid prior specification).

Although Fawcett and Walshaw (2006a) demonstrate the ease with which more complex hierarchical models can be fitted within the Bayesian framework, they do not account for any spatial structure; that is, in the model hierarchy outlined above, sites are exchangeable, an over-simplification which can be addressed by adopting a parametric form to govern the spatial dependence between extremes observed at multiple sites within a region. To this end, Davison *et al.* (2012) consider using Gaussian processes (after suitable marginal transformations), with standard correlation functions from the geostatistics literature (e.g. Diggle and Ribeiro, 2007) to represent the decay in dependence between extremes at a pair of sites with distance. Within the Bayesian context, they also consider latent variable models for rainfall extremes observed at a network of sites across a region in Switzerland, using the co-ordinates of these sites as covariates to allow interpolation of extremes at locations for which no rainfall measurements were made. On a completely continuous scale, this allows the production of 'heat maps', wherein estimated return levels can be displayed smoothly for all points within a region simultaneously. Davison *et al.* (2012) also consider *max-stable models* for spatial dependence, making use of the multivariate extension of Equation (4) and the various models for extremal dependence discussed. Currently, spatial models are a hot topic of research in the field of extremes, the implementation of which might be accessible to practitioners through the development of R packages such as `CompRandFld` (Padoan and Bevilacqua, 2013).

3.3 Predictive inference: simulation study

Throughout this Section we have demonstrated the natural extension of Bayesian inference to prediction. In particular, we have discussed the potential appeal of the predictive return level to practitioners; inference on this quantity provides a design parameter estimate with uncertainty in parameter estimation and future observations 'built in'. We now compare the sampling properties of $z_{r,\mathsf{pred}}$ to those of two commonly-used point estimates from the posterior distribution of z_r through a simulation study. Following the Bayesian analyses of wind speed extremes at Bradfield detailed in this Section, we simulate large 'master' datasets from the seasonal piecewise model (see Section 3.1) and the hierarchical model (see Section 3.2). Specifically, we use $(\bar{\sigma}_m, \bar{\xi}_m, \bar{\alpha}_m)$, the posterior means of the GPD parameters and logistic dependence parameter in the seasonal piecewise model, to simulate 10,000 wind speed extremes in each month m, $m = 1, \ldots, 12$; for the hierarchical model, we use $(\bar{\sigma}_{m,s}, \bar{\xi}_{m,s}, \bar{\alpha}_s)$ for each month m and site s, $m, s = 1, \ldots, 12$. Simulating 10,000 extremes in each month gives around 30 times as many simulated extremes as we have actual observed extremes at Bradfield. Large MCMC runs are then applied to these master datasets to obtain estimates of predictive return levels at Bradfield, these estimates being treated as the true values of $z_{r,\mathsf{pred}}$. Specifically, Equation (13) is solved for $z_{r,\mathsf{pred}}$ using, for example, $\psi^{[j]} = (\sigma_m, \xi_m, \theta_m)^{[j]}$ in the seasonal piecewise model, where θ_m is obtained from the posterior draw for α_m via Equation (9) and $j = 1, \ldots, 10^7$ after the removal of burn-in. Similarly, the means of the posterior draws for z_r from these large MCMC runs, obtained by solving Equation (8) for $x = z_r^{[j]}$ using $\psi^{[j]}$, $j = 1, \ldots, 10^7$, are taken to be the true posterior means for z_r, which we label as $z_{r,\mathsf{mean}}$. We also obtain $z_{r,\mathsf{upper}}$, the 97.5% empirical quantile of $z_r^{[j]}$, $j = 1, \ldots, 10^7$ (i.e. the upper endpoint of the 95% credible interval for z_r, often used as a design parameter in practice).

We simulate N years of wind speed extremes from each of the seasonal piecewise and hierarchical models, using $(\bar{\sigma}_m, \bar{\xi}_m, \bar{\alpha}_m)$ and $(\bar{\sigma}_{m,s}, \bar{\xi}_{m,s}, \bar{\alpha}_s)$ respectively and with the same number of simulated extremes as were observed at Bradfield (and the other sites in the hierarchical model). We then find $P_{r,\mathsf{pred}}$, $P_{r,\mathsf{mean}}$ and $P_{r,\mathsf{upper}}$ – the proportion of years in which the maximum simulated extreme exceeds $z_{r,\mathsf{pred}}$, $z_{r,\mathsf{mean}}$ and $z_{r,\mathsf{upper}}$ (respectively). This exercise is repeated L times in order to assess the variability in our estimates of these proportions. We use $N = 10,000$ and $L = 1000$. We also repeat the entire simulation procedure for other strengths of extremal dependence, and for other dependence models. For example, for the Bradfield wind speed extremes most $\bar{\alpha}_m$ were around 0.3; we also consider $\bar{\alpha}_m = 0.5$ and $\bar{\alpha}_m = 0.75$. We also consider the case of asymptotic independence through AR(1) processes with varying strengths of serial correlation, as well as other marginal shape parameters $\bar{\xi}_m$ to assess the performance of each return level estimate for different tail behaviours.

Table 6 summarises one arm of the simulation study, showing sampling properties for the different exceedance proportions for the seasonal piecewise model using $(\bar{\sigma}_m, \bar{\xi}_m, \bar{\alpha}_m)$ from the original fits to the Bradfield wind speed extremes as discussed in Section 3.1. Although not shown here, similar

		Sampling distribution		
		Mean	St. Dev.	95% CI
$r^{-1} = 10\%$	$P_{10,\mathsf{mean}}\%$	18.256	0.566	(17.126, 19.282)
	$P_{10,\mathsf{upper}}\%$	5.916	0.902	(4.223, 7.455)
	$P_{10,\mathsf{pred}}\%$	6.395	0.328	(5.815, 6.981)
$r^{-1} = 2\%$	$P_{50,\mathsf{mean}}\%$	3.386	0.258	(2.959, 3.911)
	$P_{50,\mathsf{upper}}\%$	0.240	0.172	(0.000, 0.578)
	$P_{50,\mathsf{pred}}\%$	0.239	0.069	(0.110, 0.371)
$r^{-1} = 0.5\%$	$P_{200,\mathsf{mean}}\%$	0.806	0.126	(0.570, 1.031)
	$P_{200,\mathsf{upper}}\%$	0.004	0.015	(0.000, 0.025)
	$P_{200,\mathsf{pred}}\%$	0.003	0.008	(0.000, 0.020)

Table 6 Sampling distribution summaries for $P_{r,\mathsf{mean}}$, $P_{r,\mathsf{upper}}$ and $P_{r,\mathsf{pred}}$ using $L = 1000$ repeated simulations of $N = 10,000$ years of threshold exceedances from the seasonal piecewise model obtained from fits to the Bradfield wind speed data.

findings were obtained for different $\bar{\alpha}_m$ and $\bar{\xi}_m$, and for simulations based on the hierarchical model (although, owing to the sharing of information across different sites and seasons, sampling variability was substantially reduced here); results using AR(1) processes for the dependence structure bore similar findings. The table shows results for $r = 10$, 50 and 200 years, although results for other return periods were also examined. We make several observations:

- $z_{r,\mathsf{mean}}$ consistently leads to significant over-estimates of r^{-1} (i.e. the sampling distribution means for $P_{r,\mathsf{mean}}$ are higher than the intended exceedance probabilities r^{-1}, and the 95% confidence interval lower bounds from these distributions always exceed r^{-1}). This suggests that using the posterior mean could result in substantial under-protection.
- The predictive return levels $z_{r,\mathsf{pred}}$ consistently lead to significant *under*-estimates of the intended exceedance probabilities. However, this is to be expected: these quantities have taken into account any variability in the estimates of marginal and dependence parameters, as well as uncertainty in future observations. Thus, we would expect $z_{r,\mathsf{pred}} > z_{r,\mathsf{mean}}$, leading to exceedance probabilities which are possibly smaller than r^{-1}. In practice, this could lead to over-protection. However, this might be on a par with the common practice of designing to the upper-endpoint of the 95% confidence interval for z_r (see next point), but with uncertainty in future observations also included.
- In all cases, there appears to be no significant difference in the exceedance proportions resulting from $z_{r,\mathsf{pred}}$ and $z_{r,\mathsf{upper}}$, although the sampling distribution means are, in most cases, smaller for $z_{r,\mathsf{pred}}$; see previous point.

Our simulations show that none of the return level estimators achieve their stated exceedance probabilities of r^{-1}. Although this should be expected of $z_{r,\mathsf{upper}}$ and $z_{r,\mathsf{pred}}$, the fact that $z_{r,\mathsf{mean}}$ gives consistently over-estimated values for these exceedance probabilities indicates that this posterior summary might be inadequate in any practical application. As expected, $z_{r,\mathsf{upper}}$ and $z_{r,\mathsf{pred}}$ give consistently smaller estimates of these exceedance probabilities. However, as a single number summary, both at least have uncertainty in parameter estimation built in, $z_{r,\mathsf{pred}}$ also allowing for randomness in future observations.

4 Conclusions and recommendations

We have presented a summary of the current state of play with regard to the methodology for return level estimation, and here we provide some conclusions and some recommendations for practitioners. Clearly block maxima methods are wasteful of data, and if return level estimation is the priority, they should only be considered as a serious option if block maxima are the only data available, or if other aspects of the model being implemented make it so complex that the extra structure involved with imposing threshold selection of extremes is considered a step too far. As an example, Atyeo and Walshaw (2012) take this view.

Generally threshold methods should be preferred, as they are less wasteful of data. However, given this, a key recommendation is that the traditional POT approach is discarded. In addition to being wasteful of data, the sub-asymptotic theory of this approach indicates that estimates of parameters are biased (Eastoe and Tawn, 2012) backing up empirical findings by Fawcett and Walshaw (2007). The recommended alternative is to use *all* exceedances, through careful estimation of the extremal index. On the basis of this work we recommend using one of the non-parametric intervals estimators proposed by Ferro and Segers (2003) or Süveges (2007). Our recommendation for assessing the uncertainty associated with return levels is to produce confidence intervals using a block bootstrap procedure, as described fully in Fawcett and Walshaw (2012). Alternatively, if one wishes to take a more theoretical approach than that based on estimation of the extremal index, then the sub-asymptotic behaviour of cluster peaks is derived in terms of a combination of terms based on the marginal and dependence behaviour of all exceedances respectively, giving rise to an appropriate model (Eastoe and Tawn, 2012).

The recommendations thus far are aimed at those wishing to take a frequentist approach to inference. However the authors would favour a Bayesian approach, and would recommend this to practitioners wherever their philosophical approach to inference, and their willingness to get involved with the computational issues, permit! Inference is bound to be improved by the incorporation of useful prior information, and this is almost always available in one form or another. This could be through genuine elicitation of expert beliefs, but more commonly the Bayesian approach allows for the incorporation of information from other studies, or

from other locations being considered in the same study, thereby providing a very natural route to sharing information and thereby improving estimation precision. Thus the Bayesian approach is a natural way to consider spatial or hierarchical models for extremal behaviour at multiple sites (of course, such models could be estimated in the frequentist setting; however, we believe MCMC techniques within the Bayesian framework provide a much more convenient route to inference here). Estimation uncertainty is now naturally represented in the posterior distributions of all quantities of interest, including return levels, and this information is easily extracted from any sampling scheme used for inference. Finally, there has been a long-standing clash between frequentist statisticians and many practitioners in the interpretation of return levels. In our experience, practitioners often take the view that a return level estimate, in itself a statement about probabilities, should not then need to be accompanied by an estimate of uncertainty as to its value. The Bayesian approach, unlike that of the frequentists, is fully supportive of this view. The *posterior predictive value* for a return level does exactly what is required by such practitioners, in that all uncertainty about parameter estimation (and randomness in future observations) has been integrated out in the provision of this prediction, which is then correctly interpreted as a probability statement which does not (and should not) be accompanied by an assessment of uncertainty. Of course uncertainty about the model itself is always present, but both frequentist and Bayesian perspectives are always conditioning on the fitted model being correct when presenting results, while acknowledging that this is inevitably an approximation to the truth.

Although this paper is part review / survey in nature, so vast is the literature on return level estimation – both in Statistics journals and journals of a more applied nature – that the review element of this article is not exhaustive. Indeed, readers need look no further than the SERRA journal itself to find many articles relating to the problem of return level estimation in various environmental applications. For example, papers by Shiau (2003), Xu *et al.* (2010), Galiatsatou and Prinos (2011), Vanem (2011) and Van der Vyver (2015) all tackle the issue of return level or return period estimation in a variety of contexts, most of which use methods similar to those presented in Section 1. Serinaldi (2015) also provides a very interesting SERRA communiqué on return period estimation, relevant to the work in this paper. However it is our belief that, given the compelling case for the use of Bayesian methods presented here and elsewhere, it is surprising that such methods have not yet become commonplace in practice.

To sum up then, we recommend using a method which makes use of all threshold exceedances wherever possible, and we believe that a Bayesian approach is preferable where this is feasible. Of course all models need to have a sensible (often pragmatic) approach to seasonal variation built in, and all of the modelling approaches we have described are amenable to being extended to incorporate covariate effects, including temporal trends, in the parameter values and hence the return levels. We believe the methods we propose for handling temporal dependence allow all threshold excesses to be pressed into use in a fairly simple way. Further, working within the Bayesian framework allows the estimation of the predictive return level – a quantity which lends itself to easy communication with practitioners having, as it does, all sources of uncertainty built-in. The methods we outline are robust and versatile, and could be easily applied to most environmental variables.

Acknowledgements We would like to thank three referees, and the Associate Editor, for their extremely helpful comments.

References

[1] Ancona-Navarrete, M.A. and Tawn, J.A. (2000). A Comparison of Methods for Estimating the Extremal Index. *Extremes*, **3**, pp. 5–38.

[2] Atyeo, J. and Walshaw, D. (2012). A region-based hierarchical model for extreme rainfall over the UK, incorporating spatial dependence and temporal trend. *Environmetrics*, **23**, 6, pp. 509–521.

[3] Beirlant, J., Goegebeur, J., Teugels, J., Segers, J., De Waal, D. and Ferro, C. (2004). *Statistics of Extremes*. Wiley, New York.

[4] Bottolo, P., Consonni, G., Dellaportos, P., Lijoi, A. (2003). Bayesian analysis of extreme values by mixture modelling. *Extremes*, **6**, pp. 25–48.

[5] Brooks, S.P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, **7**, 4, pp. 434–455.

[6] Cabras, S. (2013). Default priors based on pseudo-likelihoods for the Poisson-GPD model. In N. Torelli, F. Pesarin and A. Bar-Hens (Eds.), *Advances in Theoretical and Applied Statistics*, Studies in Theoretical and Applied Statistics, pp. 3–12. Springer, Berlin.

[7] Chavez-Demoulin, V. and Davison, A. (2005). General additive modelling of sample extremes. *J. R. Statist. Soc., C*, **54**, pp. 207–222.

[8] Coles, S.G. (2001). *An introduction to statistical modeling of extreme values*. Springer, London.

[9] Coles, S.G. and Tawn, J.A. (1991). Modelling Extreme Multivariate Events. *J. R. Statist. Soc., B*, **53**, pp. 377–392.

[10] Coles, S.G. and Powell, E.A. (1996). Bayesian methods in extreme value modelling: a review and new developments. *International Statistical Review*, **64**, 1, pp. 119–136.

[11] Coles, S.G. and Tawn, J.A. (1996). A Bayesian Analysis of Extreme Rainfall Data. *J. R. Statist. Soc., C*, **45**, pp. 463–478.

[12] Coles, S.G., Heffernan, J.E. and Tawn, J.A. (1999). Dependence Measures for Extreme Value Analyses. *Extremes*, **2**, pp. 339–365.

[13] Coles, S.G. and Tawn, J.A. (2005). Bayesian modelling of extreme surges on the UK east coast. *Phil. Trans, Roy. Soc. A: Mathematical, Physical and Engineering Sciences*, 363, pp. 1387–1406.

[14] Davison, A.C. and Smith, R.L. (1990). Models for Exceedances over High Thresholds (with discussion). *J. R. Statist. Soc., B*, **52**, pp. 393–442.

[15] Davison, A.C., Padoan, S.A. and Ribatet, M. (2012). Statistical Modeling of Spatial Extremes. *Stat. Sci.*, **27**, pp. 161–186.

[16] Diggle, P.J. and Ribeiro, P.J. (2007). *Model-based Geostatistics*, Springer, New York.

[17] Eastoe, E.F. (2009). A hierarchical model for non-stationary multivariate extremes: a case study of surface-level ozone and NOx data in the UK. *Environmetrics*, **20**, pp. 428–444.
[18] Eastoe, E. F. and Tawn, J.A. (2012). Modelling the distribution for the cluster maxima of exceedances of sub-asymptotic thresholds. *Biometrika*, 99, **1**, pp. 43–55.
[19] Efron, B. (1987). Better bootstrap confidence intervals. *J. Amer. Stat. Ass.*, **82**, pp. 171–185.
[20] Eugenia Castellanos, M. and Cabras, S. (2007). A default Bayesian procedure for the generalized Pareto distribution. *J. Stat. Plan. Inf*, **137**, 2, pp. 473–483.
[21] Fawcett, L. (2005). Statistical Methodology for the Estimation of Environmental Extremes. *PhD Thesis*, Newcastle University, Newcastle-upon-Tyne.
[22] Fawcett, L. and Walshaw, D. (2006). Markov Chain Models for Extreme Wind Speeds. Environmetrics, **17**, 8, pp. 795–809.
[23] Fawcett, L. and Walshaw, D. (2006a). A hierarchical model for extreme wind speeds. *J. Roy. Statist. Soc., C*, **55**, 5, pp. 631–646.
[24] Fawcett, L. and Walshaw, D. (2007). Improved Estimation for Temporally Clustered Extremes. *Environmetrics*, **18**, 2, pp. 173–188.
[25] Fawcett L, Walshaw D. (2008). Bayesian inference for clustered extremes. *Extremes*, **11**, pp. 217–233.
[26] Fawcett, L. and Walshaw, D. (2012). Estimating return levels from serially dependent extremes. *Environmetrics*, 23, **3**, pp. 272–283.
[27] Ferro, C.A.T. and Segers, J. (2003). Inference for clusters of extreme values. *J. R. Statist. Soc., B*, **65**, pp. 545–556.
[28] Galiatsatou, P. and Prinos, P. (2011). Modeling non-stationary extreme waves using a point process approach and wavelets. *Stoch. Environ. Res. Risk Assess.*, **25**, 2, pp. 165–183.
[29] Gomes, M.I., (1993). On the estimation of parameters of rare events in environmental time series. In *Statistics for the environment 2: Water Related Issues* (V. Barnett and K.F. Turkman), pp. 225–241.
[30] Ho, K.W. (2010). A matching prior for extreme quantile estimation of the generalized Pareto distribution. *J. Stat. Plan. Inf.*, **140**, 6, pp. 1513–1518.
[31] Hsing, T. (1993). Extremal index estimation for a weakly dependent stationary sequence. *Ann. Statist.*, **21**, pp. 2043–2071.
[32] Jonathan, P., Ewans, K.C. and Randell, D. (2014). Non-stationary conditional extremes of northern North Sea storm characteristics. *Environmetrics*, **25**, 3, pp. 172–188.
[33] Jonathan, P. and Ewans, K.C. (2011). Modelling the seasonality of extreme waves in the Gulf of Mexico. *ASME J. Offshore Mech. Arct. Eng.*, 133:021104.
[34] Leadbetter, M.R., Lindgren, G. and Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Series*. Springer-Verlag, New York.
[35] Leadbetter, M.R. and Rootzén, H. (1988). Extremal theory for stochastic processes. *Ann. Probab.*, **16**, pp. 431–476.
[36] Northrop, P. (2012). Semiparametric estimation of the extremal index using block maxima. Technical report number 318, Department of Statistical Science, University College London.
[37] Northrop, P. and Attalides, N. (2014). Posterior propriety in objective Bayesian extreme value analyses. Departmental research report 323, University College London.
[38] Padoan, S.A. and Bevilacqua, M. (2013). *CompRandFld: Composite-likelihood based Analysis of Random Fields*, R package version 1.0.3.
[39] Pickands, J. (1981). Multivariate extreme value distributions, in *Proceedings of the 43rd session of the International Statistics Institute, Vol. 2*, *Bull. Inst. Internat. Statist.* **49**, pp. 859–878.
[40] Roberts, G.O., Gelman, A. and Gilks, W.R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, **7**, 1, pp. 110–120.
[41] Sang, H. and Gelfand, A.E. (2009). Hierarchical Modeling for Extreme Values Observed over space and time. *Environmental and Ecological Statistics*, **16**, pp. 407–426.
[42] Sang, H. and Gelfand, A.E. (2010). Continuous Spatial Process Models for Extreme Values. *Journal of Agricultural, Biological and Environmental Statistics*, **15**, pp. 49–65.
[43] Scarrott, C. and MacDonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT - Statistical Journal*. **10**, 1, pp. 33–60.
[44] Schlather, M. (2002). Models for stationary max-stable random fields, *Extremes*, **5**, pp. 33–44.
[45] erinaldi, F. (2015). Dismissing return periods! *Stoch. Environ. Res. Risk Assess.*, **29**, 4, pp. 1179–1189.
[46] Shiau, J.T. (2003). Return period of bivariate distributed extreme hydrological events. *Stoch. Environ. Res. Risk Assess.*, **17**, 1-2, pp. 42–57.
[47] Smith, A.F.M. and Roberts, G.O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Statist. Soc., B*, **55** pp. 3–23.
[48] Smith, E.L. and Walshaw, D. (2003). Modelling Bivariate Extremes in a Region. *Bayesian Statistics*, **7**, pp. 681–690.
[49] Smith, R.L. (1992). The Extremal Index for a Markov Chain. *J. Appl. Prob*, **29**, pp. 37–45.
[50] Smith, R.L. (1999). Bayesian and frequentist approaches to parametric predictive inference (with discussion). In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (eds.) *Bayesian Statistics*, **6**, pp. 589–223.
[51] Smith , R.L. and Weissman, I. (1994). Estimating the Extremal Index. *J. R. Statist. Soc., B*, **56**, pp. 515–528.
[52] Smith, R.L., Tawn, J.A. and Coles, S.G. (1997). Markov chain models for threshold exceedances. *Biometrika*, **84**, pp. 249–268.
[53] Smith, R.L., Goodman, D.J. (2000). Bayesian risk analysis. In: Embrechts, P. (ed.) *Extremes and Integrated Risk Management*, pp. 235–251. Risk Books, London.
[54] Stepheson, A. and Ribatet, M. (2014). *evdbayes: Bayesian Analysis in Extreme Value Theory*, R package version 1.8.0.
[55] Stephenson, A. and Tawn, J. A. (2004). Inference for extremes: accounting for the three extremal types. *Extremes*, **7**, 4, pp. 291–307.
[56] Süveges, M. (2007). Likelihood estimation of the extremal index. *Extremes*, **10**, pp. 41–55.
[57] Süveges, M. and Davison, A.C. (2010). Model misspecification in peaks over threshold analysis. *Ann. Appl. Statist.*, **4**, pp. 203–221.
[58] Van der Vyver, H. (2015). On the estimation of continuous 24-h precipitation maxima. *Stoch. Environ. Res. Risk Assess.*, **29**, 3, pp. 653–663.
[59] Vanem, E. (2011). Long-term time-dependent stochastic modelling of extreme wave. *Stoch. Environ. Res. Risk Assess.*, **25**, pp. 185–209.
[60] Walshaw, D. (1991). Statistical Analysis of Extreme Wind Speeds. *PhD Thesis*. University of Sheffield, Sheffield.
[61] Walshaw, D. (1994). Getting the Most From Your Extreme Wind Data: A Step by Step Guide. *J. Res. Natl. Inst. Stand. Technol.*, **99**, pp. 399–411.
[62] Xu, Y., Booij, M.J. and Tang, Y. (2010). Uncertainty analysis in statistical modeling of extreme hydrological events. *Stoch. Environ. Res. Risk Assess.*, **24**, 5, pp. 567–578.

Appendix

A1: Extremal index estimators

Cluster size estimators

- THE RUNS ESTIMATOR: $\hat{\theta} = (\text{mean cluster size})^{-1}$, using cluster termination interval κ to identify clusters (see Section 1.2).

- THE BLOCKS ESTIMATOR: As for the runs estimator, but where blocks of length τ are considered clusters if there is at least one threshold exceedance within the block.

Maxima methods

- GOMES' ESTIMATOR: Obtain $(\hat{\mu}_\theta, \hat{\varsigma}_\theta, \hat{\xi}_\theta)$ for the GEV applied to block maxima $\{M_\tau\}$ with block length τ. Find also $(\hat{\mu}, \hat{\varsigma}, \hat{\xi})$ from block maxima $\{\bar{M}_\tau\}$, obtained from an independent series after randomisation of the original series. Then

  $\hat{\theta} = (\hat{\varsigma}/\hat{\varsigma}_\theta)^{-1/\hat{\xi}}$, where
  $\hat{\xi} = (\hat{\varsigma} - \hat{\varsigma}_\theta)/(\hat{\mu} - \hat{\mu}_\theta)$. *[Gomes (1993)]*

- NORTHROP'S ESTIMATOR:

  $$\hat{\theta} = -1/\overline{\log V},$$

  with $\overline{\log V} = \sum_{i=1}^{n} \log V_i/n$, V_i being a random sample from a $Beta(\theta, 1)$ distribution. *[Northrop (2012)]*

Intervals estimators

- FERRO AND SEGERS' ESTIMATOR:

  $$\hat{\theta} = \min\left\{1, \sum_{i=1}^{J-1}(T_i - a)^2/(J-1)\sum_{i=1}^{J-1}(T_i - b)(T_i - c)\right\},$$

  where $T_i = S_{i+1} - S_i$, $i = 1, \ldots, J-1$ are the times between J threshold exceedances; $a = b = c = 0$ if $\max(T_i) \leq 2$; otherwise, $a = b = 1, c = 2$. *[Ferro & Segers (2003)]*
- SÜVEGES' MLE: Maximum likelihood estimator based on an extension of the work in Ferro & Segers (2003). The likelihood for $U_i = T_i - 1$, $i = 1, \ldots, J-1$, is maximised to obtain a closed-form expression for $\hat{\theta}$. *[Süveges (2007)]*
- SÜVEGES' IWLS: Iterative weighted least squares estimator based on the normalised gaps between clusters. *[Süveges (2007)]*
- K-GAPS ESTIMATOR: An extension of Süveges' MLE, shown to have reduced bias and RMSE (given an optimal choice of tuning parameter K). *[Süveges & Davison (2010)]*

A2: Bayesian sampling in the hierarchical model

For the hierarchical model outlined in Section 3.2, we have

$$\eta_{m,s} = \gamma_\eta^{(m)} + \varepsilon_\eta^{(s)},$$
$$\xi_{m,s} = \gamma_\xi^{(m)} + \varepsilon_\xi^{(s)} \quad \text{and}$$
$$\alpha_s = \varepsilon_\alpha^{(s)}$$

for the GPD (log) scale and shape, and the logistic dependence parameters (respectively). All random effects for $\eta_{m,s}$ and $\xi_{m,s}$ are assumed to be normally distributed:

$$\gamma_\eta \sim N(0, \tau_\eta^{-1}) \qquad \text{and}$$
$$\gamma_\xi^{(m)} \sim N(0, \tau_\xi^{-1}), \qquad m = 1, \ldots, 12,$$

for seasonal effects, and

$$\varepsilon_\eta^{(s)} \sim N(a_\eta, \zeta_\eta^{-1}) \qquad \text{and}$$
$$\varepsilon_\xi^{(s)} \sim N(a_\xi, \zeta_\xi^{-1}), \qquad s = 1, \ldots, 12,$$

for site effects. We choose the mean of the normal distribution of the seasonal effects to be fixed at zero to avoid over-parameterisation and problems of identifiability; however, we could equally have fixed the mean for the distribution of site effects to achieve this. Since the logistic dependence parameter α must lie between 0 and 1, we draw the site effect for α from a uniform distribution, and so

$$\varepsilon_\alpha^{(s)} \sim U(0, 1).$$

The final layer of the model is to specify prior distributions for the random effect distribution parameters. Here, we have chosen largely non-informative priors, adopting conjugacy wherever possible to simplify computations. Thus,

$$a_\eta \sim N(b_\eta, c_\eta), \quad a_\xi \sim N(b_\xi, c_\xi),$$
$$\tau_\eta \sim Ga(d_\eta, e_\eta), \quad \tau_\xi \sim Ga(d_\xi, e_\xi), \qquad \text{and}$$
$$\zeta_\eta \sim Ga(f_\eta, g_\eta), \quad \zeta_\xi \sim Ga(f_\xi, g_\xi),$$

with a suitable specification of hyper-parameters. The MCMC algorithm employed is Metropolis within Gibbs, i.e. we update each component singly using a Gibbs sampler where the conjugacy allows straightforward sampling from the full conditionals, and a Metropolis step elsewhere. The full conditionals for the Gibbs sampling are:

$$a_\cdot | \ldots \sim N\left(\frac{b_\cdot c_\cdot + \zeta_\cdot \sum \varepsilon_\cdot^{(s)}}{c_\cdot + n_s \zeta_\cdot}, c_\cdot + n_s \zeta_\cdot\right),$$

$$\zeta_\cdot | \ldots \sim Ga\left(f_\cdot + \frac{n_s}{2}, g_\cdot + \frac{1}{2}\sum(\varepsilon_\cdot^{(s)} - a_\cdot)^2\right)$$

and

$$\tau_\cdot | \ldots \sim Ga\left(d_\cdot + \frac{n_m}{2}, e_\cdot + \frac{1}{2}\sum(\gamma_\cdot^{(m)})^2\right),$$

where n_m = number of months = 12 and n_s = number of sites = 12, and here the notation $\zeta_\cdot$, for example, is used generically to denote either ζ_η or ζ_ξ. The complexity of the likelihood derived from the GPD means that conjugacy is unattainable for the random effect parameters, and a Metropolis step is used to update each of these.

In the absence of expert prior knowledge, prior parameters were chosen to give a highly non-informative specification:

$$b_\cdot = 0, c_\cdot = 10^{-6}, d_\cdot = e_\cdot = f_\cdot = g_\cdot = 10^{-2}.$$